Effort in Low-Stakes Assessments: What Does It Take to Perform as Well as in a High-Stakes Setting?

Educational and Psychological

Measurement

2016, Vol. 76(6) 1045–1058

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416634789

epm.sagepub.com

Yigal Attali¹

Abstract

Performance of students in low-stakes testing situations has been a concern and focus of recent research. However, researchers who have examined the effect of stakes on performance have not been able to compare low-stakes performance to truly high-stakes performance of the same students. Results of such a comparison are reported in this article. GRE test takers volunteered to take an additional low-stakes test, of either verbal or quantitative reasoning as part of a research study immediately following their operational high-stakes test. Analyses of performance under the high- and low-stakes situations revealed that the level of effort in the low-stakes situation (as measured by the amount of time on task) strongly predicted the stakes effect on performance (difference between test scores in low- and high-stakes situations). Moreover, the stakes effect virtually disappeared for participants who spent at least one-third of the allotted time in the low-stakes situation. For this group of test takers (more than 80% of the total sample), the correlations between the low- and high-stakes scores approached the upper bound possible considering the reliability of the test.

Keywords

low-stakes assessment, motivation, filtering

As part of accountability policies in education systems, students are required to participate in many assessments that have little or no consequences for them. Without consequences for performance, many students will not give their best effort to such

¹Educational Testing Service, Princeton, NJ, USA

Corresponding Author:

Yigal Attali, Educational Testing Service, Rosedale Road, MS-10-R, Princeton, NJ 08541, USA.

Email: yattali@ets.org

low-stakes tests. If less-motivated students score below their actual proficiency level, test-taking motivation would be a possible source of bias (Wainer, 1993) and hence a threat to the validity of score interpretation and use (Messick, 1995). The validity implications of lower motivation in low-stakes test are not restricted to the individual level. If different groups of students differ in level of motivation, interpretation of group differences would be compromised. Wainer noted that in an international assessment administration in Korea, the participating students (although chosen at random) were told that they were representing the honor of their school and their country, and could therefore have had higher motivation than typical American students to succeed on the test.

Past research supports the assumption that tests with no personal consequences, that is, low-stakes tests, are associated with a decrease in motivation and performance (Wise & DeMars, 2005; Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995). Wise and DeMars (2005) reviewed 12 studies that included 25 comparisons between the performance of motivated and less-motivated groups of test takers. In most comparisons, the performance of the motivated group was significantly higher than the less-motivated group, and the average d effect size was around 0.6. The manipulations used to motivate test takers varied. In most cases, the students were told the scores would count toward course or school grades. In a recent study, Liu, Bridgeman, and Adler (2012) found that increasing the personal stakes of a test by telling test takers that their scores could be released to faculty in their college or potential employers significantly increased test performance, with a d effect size of 0.41.

On the other hand, studies that used monetary incentives to motivate students in tests have generally found weak effects on performance. O'Neil, Sugrue, and Baker (1996) offered 8th- and 12th-grade students taking the National Assessment of Educational Progress (NAEP) \$1 for each item they answered correctly, and their performance was compared with control groups that received standard NAEP instructions. A significant effect was found only for a subsample of 8th graders (those who correctly remembered their treatment condition), and only on the easy items. More recently, this study was replicated (O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005) with \$10 per item, but again no effect on scores was found. Baumert and Demmrich (2001) offered one group of students a flat payment if they answered more items than expected based on their school grade, but failed to find an effect on scores. Finally, Braun, Kirsch, and Yamamoto (2011) administered a NAEP assessment to 12th-grade students and found weak effects (d effect size of 0.08 to 0.25) for monetary incentives (either a fixed amount of \$20 or variable according to performance) compared with a control group.

These results demonstrate that monetary incentives do not seem to motivate students to perform well in tests. In addition, such extrinsic rewards may even undermine students' motivation in the long run. Although early research has demonstrated the power of extrinsic rewards in controlling behavior (Skinner, 1953), later research has shown that for intrinsically motivated activities (and academic performance is

one such activity), tangible rewards may undermine this intrinsic motivation (Deci, Koestner, & Ryan, 1999). That is, rewarding students for performing well in an educational setting may be counterproductive in the long run.

In summary, previous experimental research seems to suggest that the most effective method for raising student motivation in low-stakes tests is to manipulate the personal stakes for test takers, most commonly by telling them that test scores would count toward course grades (Sundre, 1999; Wolf & Smith, 1995).

In real-world low-stakes contexts, it is usually not possible to manipulate test stakes; nor is it possible to manipulate the information about stakes that is communicated to test takers. For these situations, it is important to understand the nature and magnitude of the threat posed by lack of test taker effort. Most studies (e.g., Sundre, 1999; Wolf & Smith, 1995) administer posttest self-report questionnaires on effort and motivation. However, self-report measures of effort are potentially vulnerable to bias through motivational processes (Wise, 2006). For example, some test takers may be reluctant to admit they did not make good effort. Alternatively, test takers who feel they did not do well on the test may underreport their effort. An alternative objective measure of test taker effort based on item response times in computer based tests was suggested by Wise and Kong (2005). They suggested that a prevalence of very short response times can be indicative of low effort since under these conditions the student does not even have time to fully consider the item. Wise and Kong defined a *response time effort* person index as the proportion of short rapid-guessing responses in a test and found that this index had high internal consistency and exhibited convergent validity through positive correlations with self-report measures of effort.

In several studies, the impact of *motivation filtering* on test results was also explored. The rationale for motivation filtering is that when test takers do not give good effort in a test, their test scores are not a valid indicator of their true proficiency level. By deleting the data from these test takers, the data from the remaining test takers should provide a more accurate estimate of the actual mean proficiency of the group. Based either on self-report measures (Sundre & Wise, 2003; Wise & DeMars, 2005) or on response time effort measures (Wise & DeMars, 2010; Wise & Kong, 2005) research has shown that when test takers assumed to have shown low effort are excluded from test analyses, average test scores increase, variability of test scores decreases, and the correlations between test scores and external variables expected to correlate with test scores (i.e., convergent validity evidence) increase. This research has also shown that test taker effort is unrelated to ability, as measured with an independent high-stakes assessment. All together, these findings are consistent with the conclusion that motivation filtering removes construct-irrelevant variance caused by test taker lack-of-effort behavior.

Present Study

A limitation of past research on motivation and effort in low-stakes assessment is that it is based on between-subject comparisons. That is, motivational manipulations

were administered to different groups of students. Consequently, measures of effort and test performance were compared at the aggregate level, and only general conclusions on the effect of stakes were drawn. In contrast, not much is known about the relationship between ability, as measured under high-stakes conditions, and effort and performance under low-stakes conditions. One exception is that two studies (Liu et al., 2012; Wise & Kong, 2005) have found that ability, as measured by performance on an independent high-stakes assessment (the SAT in both of these studies) is very weakly correlated with effort under the low-stakes condition.

However, because of the between-subject design of previous studies little is known about how different levels of effort under the low-stakes conditions affect the stakes effect—the difference in performance between the high- and low-stakes situations.

Two questions arise in this context. First, is the relation between effort and the stakes effect more monotonous in nature, where an increase in measured effort generally results in a lower stakes effect, or is it the case that low effort must exceed a certain threshold in order to have a noticeable negative effect on performance? Second, can good effort lead to the elimination of the stakes effect, or is it the case that even when test takers show reasonable effort under low-stakes conditions, their test performance is still lower than under high-stakes conditions? Finally, are the answers to these questions dependent on the type of test content?

To study these questions, existing data from a GRE General Test study were used. The goal of the original study was to examine the effect of extra time on GRE test scores (Bridgeman, Cline, & Hessinger, 2004). For this purpose, GRE test takers were asked to take an extra research section, either in verbal reasoning (V) or quantitative reasoning (Q), at the end of their regular operational GRE test. They were told this research section would not have any effect on their operational scores. Their incentive was eligibility for a cash award that was promised to 100 participants (out of 30,000 participants) who would receive the highest scores on the research section, relative to their scores on the operational section. This promise for a chance to win the award associates some stakes to the research section. Nevertheless, it is clear that the stakes on the research section were much lower than on the regular operational test.

The focus of the original study was on comparing research test scores under two time limits (standard time and 1.5 times standard time). The purpose of the present study was to analyze the stakes effect, that is differences in performance between the research (low stakes) and operational (high stakes) test sections. The total time spent answering the research section was available and used as a response time effort measure (the time spent on the operational section was not available for analysis, as well as individual response times for research section items). Although there were very small differences in performance between the two time limits (Bridgeman et al., 2004), the analyses in this study focused on participants who were assigned to the standard time limit for the research section, in order to allow a direct comparison between performance in the research and operational sections.

Method

Materials

The GRE General Test is a standardized test used for admissions purposes in most graduate schools in the United States. The quantitative reasoning section measures problem-solving ability using basic concepts of arithmetic, algebra, geometry, and data analysis. The verbal reasoning section measures the ability to analyze and evaluate written material and synthesize information obtained from it, analyze relationships among component parts of sentences and recognize relationships among words and concepts. The analytical writing section measures critical thinking and analytical writing skills. The GRE General Test was (until 2011, when a revised test was launched) an item-level computer-adaptive test (CAT), meaning that items are dynamically selected for presentation to test takers from a pool of available items, based on the test taker's previous responses and test design decisions. The quantitative and verbal research section items were selected from regular full CAT pools (more than 300 items each) that did not overlap with the pools used for the operational sections. The only thing that distinguished these sections from regular sections was the timing (in the extended time limit) and a screen indicating that the research section did not count as part of the official test score. As a consequence, scoring (on a 200-800 scale) was equivalent in operational and research sections (see Mills & Steffen, 2000, for a detailed discussion of the technical aspects of the GRE CAT implementation).

Participants

A total of 80,979 test takers took the GRE test during the period the research project was conducted. Of these, 29,960 test takers (37%) volunteered to participate in the research study. The operational quantitative test scores of all (standard and 1.5 times standard time) participants ($M = 637$, $SD = 144$) were somewhat higher than nonparticipants ($M = 628$, $SD = 143$), whereas the operational verbal test scores of participants ($M = 475$, $SD = 124$) were somewhat lower than nonparticipants ($M = 482$, $SD = 123$). However, in both cases the differences were very small, 0.06 in standardized scores, indicating that participants were representative of the test taker population (see also Bridgeman et al., 2004). Analyses for this study include only participants assigned to the standard time condition, 15,201 in total.

Design

Participants were randomly assigned to one of four groups, defined by the type of research section administered (Q or V) and the time limit (standard or 1.5 times standard time). Time limits for Q were 45 minutes (standard) and 68 minutes. Time limits for V were 30 minutes (standard) and 45 minutes.

Table 1. Operational and Research Section Scores and Time Spent on Research Section for Standard Time Participants.

	<i>N</i>	Operational		Research		Time in Research	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Q	7,328	635.9	145.4	566.4	192.2	28.5	13.7
V	7,873	475.3	123.4	424.8	136.8	18.9	8.8

Note. Time is in minutes. Q = quantitative section; V = verbal section.

Analyses

To compare performance on the research and operational sections, standardized differences (based on operational standard deviations) were computed. No significance tests were computed for these differences because of the large sample sizes in this study.

To analyze the relationship between performance (on both the research and operational sections) and effort as measured by total time spent on the research section the R^2 for a quadratic least-squared regression model of performance on effort was calculated. A graphical method was also employed by fitting a local regression smoothing model of the scatterplots of scores and time spent on the research section. The smoothed curves were based on a weighted quadratic least squares regression (Cleveland, 1979). Local regression is nonparametric in the sense that the fitting technique does not require an a priori specification of the relationship between the dependent and independent variables, making it ideal for modeling complex processes for which no theoretical models exist (Cleveland & Devlin, 1988).

To analyze the effect of effort filtering, where filtering was based on the total time spent on the research section, standardized differences between research and operational scores were computed as well as correlations between research and operational scores, for both the entire sample and for filtered sample.

Results

As was explained in the participant section above, the following analyses include only participants in the standard time condition, 15,201 in total. Table 1 presents descriptive information on the operational and research scores for standard-time participants, as well as time spent (in minutes) on the research section. The table shows that the average research scores are considerably lower than the average operational scores, with standardized differences (based on operational standard deviations) of .48 for Q and .41 for V. The average time spent on the research section is 63% of the total standard time for each section (45 minutes for Q and 30 minutes for V). Notably, the standard deviations of the research scores are larger than those for the operational scores, especially for Q. The reason for this is evident from Figure 1,

Figure 1. Score histograms for quantitative (Q) and verbal (V) tests on operational and research sections.

which presents the distributions of both types of scores for each test type. For both test types, a relatively large number of test takers (8% for Q and 7% for V) received the minimum score of 200 on the research section, compared to 0.2% for the operational sections. Apart from this difference the research and operational distributions are quite similar.

It is clear that the prevalence of minimum scores on the research section indicates low effort on the research section by some participants. In order to better understand effort by participants, we turn to an analysis of time spent on the research section and its relation to performance. Figure 2 presents the distributions of time spent on the research sections. It shows that for both test types there appears to be three modes: the first around 2 minutes, the second around three-fourths of the maximum allotted time (33 minutes for Q and 21 minutes for V), and the third at the maximum allotted time.

It is safe to assume that spending just a few minutes cannot result in a high score. In fact, 90% of the test takers who received the minimum score of 200 on the research section spent less than 5 minutes on this section. However, Figure 3 presents a more complete picture of the relation between time and score through local regression smoothing (Cleveland, 1979) of the research score on time spent on the research

Figure 2. Histograms of time spent on the research section.

Figure 3. Local regression of research section score on time spent on the research section. Q = quantitative section; V = verbal section.

section. It shows a strong relation between the two measures, especially for lower levels of time spent. The R^2 for a quadratic least-squared regression model is .56 for Q and .33 for V. Note that for higher levels of time spent, and especially for the V test, the expected research score decreases as time spent increases. This suggests that among participants who put reasonable effort in the research section, higher ability test takers require less time to complete the test than lower ability test takers.

In contrast to the strong relation between effort and research scores, the relation between ability, as measured by the operational test scores, and effort on the research low-stakes test is very weak. For Q, a weak positive relation was found with a correlation of .10 ($R^2 = .01$) and the R^2 for a quadratic least-squared regression model was .03. For V, a weak negative relation was found with a correlation of $-.18$ ($R^2 = .03$) and the R^2 for a quadratic least-squared regression model was .05. These results replicate previous findings on the lack of relation between effort and ability.

To examine the effects of motivation filtering on research test scores and correlations with operational test scores, these measures were calculated with successively stricter filtering criteria on the basis of time spent on the research section. Table 2 presents the results of these analyses. The first line of the table presents the unfiltered results, with a stakes standardized effect (d) of .48 and .41 for Q and V, respectively, and correlations between the scores of .56 and .55, respectively. A 5-minute threshold (second line) results in filtering out 11% and 12% of participants, reduces the stakes effect to .18 and .17, and increases the score correlations to .80 and .76. The stakes effects continue to drop thereafter—a threshold of one-third of the time allotted (15 minutes for Q and 10 minutes for V) filters out 18% or 19% of participants, reduces the stakes effect to .06 or .09, and increases the score correlations to .90 or .80. Beyond this threshold there appear to be no more gains in the correlations between scores. In this respect, it is important to note that the reliability (internal consistency) of the Q and V operational scores is .90 for each of the sections (Educational Testing Service, 2008, p. 20). In other words, .90 is the expected correlation between scores on two operational test forms administered one after the other. Therefore, it appears that for the Q test the correlation of the filtered research scores with operational scores is as high as can be expected if the filtered research scores were actually operational scores. For the V test, this correlation is only slightly lower than what would be expected if the filtered research scores were operational scores.

The operational scores of filtered samples (such as with a threshold of one-third of the time allotted) also show very similar means and standard deviations to those of the full sample, and standard deviations of filtered operational scores are similar to the standard deviations of filtered research scores. In summary, after filtering out less than 20% of participants showing low effort, the distributions of research scores are very similar to those of operational scores for both Q and V tests.

Table 2. Response Time Effort Screening.

Threshold	Q						V					
	%	Operational		Research		d	%	Operational		Research		d
		M	SD	M	SD			M	SD	M	SD	
0	100	636	145	566	192	.48	100	475	123	425	137	.55
5	89	634	145	608	160	.18	88	471	122	450	124	.76
10	85	634	144	621	149	.87	81	470	122	459	122	.80
15	82	636	143	628	144	.90	71	467	121	461	122	.82
20	75	644	139	639	138	.03	53	456	121	454	122	.80
25	66	651	135	652	131	.91	32	440	121	439	123	.77
30	53	659	130	664	125	.90						
35	39	662	124	671	117	.89						
40	26	660	118	672	112	.86						

Note. Q = Quantitative section; V = Verbal section. Threshold for inclusion is minutes spent on research section; % = included out of total sample; r = correlation between scores on operational and research sections; d = standardized stakes effect based on operational SD.

Discussion

The findings from the analyses of the GRE study that are summarized in this article indicate that test takers can replicate their high-stakes performance in a low-stakes setting with relatively little effort. Test takers (more than 80% of all test takers) who chose to work on the test for at least one-third of the time allotted in the low-stakes setting earned scores that were almost indistinguishable from the scores they earned in a high-stakes setting. Moreover, the correlations between the low- and high-stakes scores for the filtered sample were very high, indicating that the quality of measurement in the low-stakes setting was as high as that in the high-stakes setting. Finally, results across the verbal and quantitative tests were very similar although the cognitive demands of the two tests are markedly different.

These results suggest that motivation to perform well in such tests does not necessarily, or even primarily, depend on personal stakes or extrinsic rewards, and intrinsic motivation to perform well is a powerful determinant of performance.

To be sure, some effort is required to perform well on these tests. However, the results of this study suggest a discontinuity in the amount of effort required. At least as measured by time spent on the research section, results do not support the argument that monotonically increasing effort leads to increasing performance in the low-stakes setting. Instead, a minority of test takers who invested very little time in this section seemed to be responsible for the lower performance of the entire group of participants in the research section. Removing these participants was enough to eliminate the stakes effect. In other words, sitting for a minimum amount of time is a sufficient prerequisite for good performance in these general ability assessments.

This result would be less noteworthy for a test with simple cognitive demands. For example, in a general knowledge multiple-choice test composed of short prompts (e.g., “Who is the inventor of the light bulb?”), answering these questions is based on relatively effortless retrieval of facts from memory. Therefore, it may seem reasonable to assume that test takers who take the time to read the prompts would perform well under low-stakes situations. However, in quantitative problem solving and verbal reasoning, a repeated effort is required to answer each item beyond reading the prompt.

A limitation of this study is that the order of the low- and high-stakes sections was not counterbalanced and that the low-stakes section was completed immediately following the high-stakes section. Two threats to interpretation of the results are therefore possible. The first is that participants might have learned, or matured, during the operational section in a way that positively contributed to their performance in the later research section. Although possible, this is unlikely in the present circumstances since most GRE test takers are well prepared for the test, having familiarized themselves with the test content and taken practice tests (Powers, 2012). A second threat to interpretation of the results is that fatigue from completing the operational test might have negatively affected participants’ performance in the research section. This possibility is more plausible (taking a standardized test is widely regarded as exhausting) but to the extent that this effect contributed to the present results, it would

have *increased* the stakes effect and lower the measurement quality of the research section.

Another limitation of the study is that test takers volunteered to participate, and therefore self-selection could have contributed to the results. Specifically, it is possible that test takers who were less fatigued by the operational test tended to volunteer for the research study. This would limit the generalizability of the results. Although this possibility cannot be ruled out, it is noteworthy that almost no differences between participants and nonparticipants were found in terms of operational test performance.

A related issue concerns the concept of optimal performance. Although in the present context low-stakes performance is considered as suboptimal to high-stakes performance, a separate research literature is focused on the perils of high-stakes situations. In these high-stakes situations, the desire to perform as well as possible can create performance pressure that could negatively affect performance. The term *choking under pressure* has been used to describe situations where pressure to perform well leads to lower than expected performance, given one's skill level (Baumeister, 1984; Beilock & Carr, 2001). A possible explanation for this phenomenon is that high-pressure situations harm performance by diverting individuals' attention to task-irrelevant thoughts, such as worries about the situation and its consequences (Beilock & Carr, 2001). The concept of choking under pressure is a reminder that both very high and very low motivation to succeed can lead to suboptimal performance.

In conclusion, this study contributes to the growing body of research on the relationship between test taker motivation and performance by demonstrating that test takers who invest even a minimal effort can perform in low-stakes situations as well as in a high-stakes situations.

Author's Note

The author would like to thank Brent Bridgeman and Fred Cline for kindly providing the data from their study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46, 610-620.

- Baumert, J., & Demmrich, A. (2001). Testing motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441-462.
- Beilock, S. L., & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General, 130*, 701-725.
- Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th grade NAEP reading assessment. *Teachers College Record, 113*, 2309-2344.
- Bridgeman, B., Cline, F., & Hessinger, J. (2004). Effect of extra time on verbal and quantitative GRE scores. *Applied Measurement in Education, 17*, 25-37.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829-836.
- Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83*, 596-610.
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*, 627-668.
- Educational Testing Service. (2008). *GRE guide to the use of scores 2008-09*. Princeton, NJ: Author. Retrieved from <http://www.ets.org/gre/edupubs>
- Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*, 352-362.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.
- Mills, C. N., & Steffen, M. (2000). The GRE computerized adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Boston, MA: Kluwer.
- O'Neil, H. F., Jr., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185-208.
- O'Neil, H. F., Jr., Sugrue, B., & Baker, E. (1996). Effects of motivational interventions on the NAEP mathematics performance. *Educational Assessment, 3*, 135-157.
- Powers, D. E. (2012). *Understanding the impact of special preparation for admissions tests* (Research Report 12-05). Princeton, NJ: Educational Testing Service.
- Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Macmillan.
- Sundre, D. L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the meeting of the American Educational Research Association, Montreal, Quebec, Canada.
- Sundre, D. L., & Wise, S. L. (2003, April). *"Motivation filtering": An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, Illinois.
- Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1-21.
- Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19*, 95-114.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

- Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*, 27-41.
- Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183.
- Wolf, L. F. & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227-242.
- Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341-351.